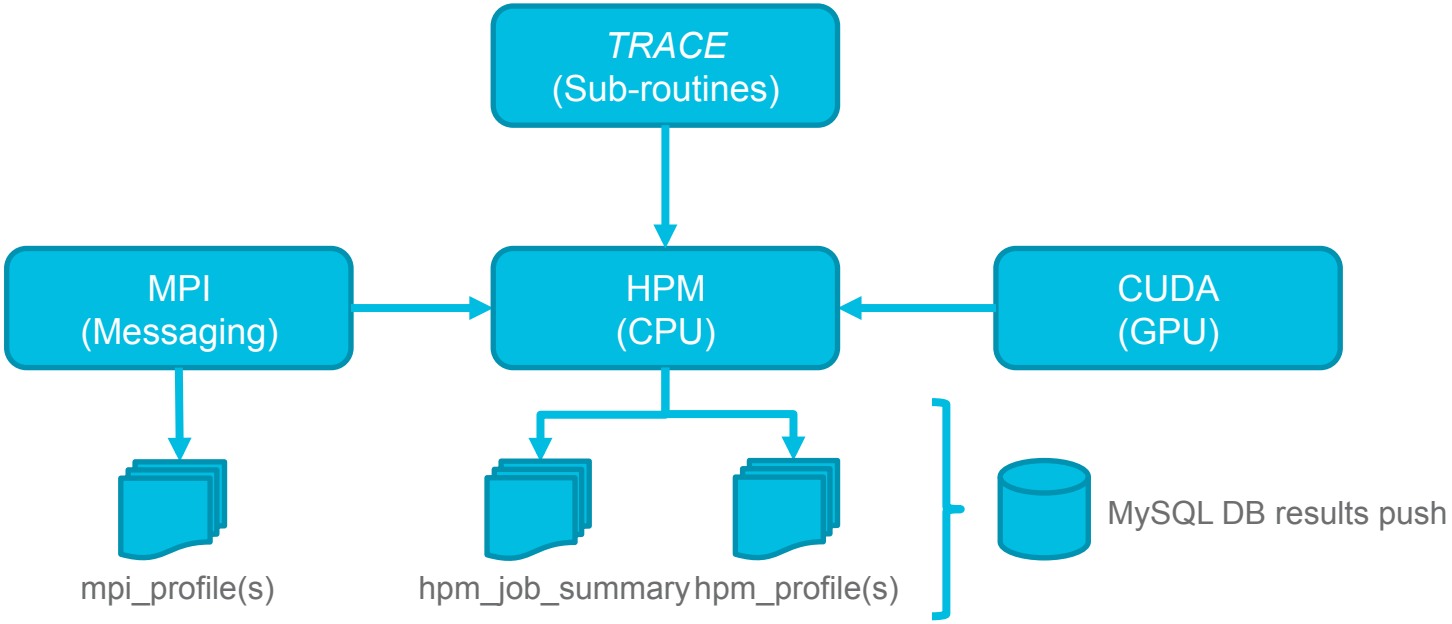
# Next-gen profiling-infrastructure for supercomputers based on hybrid nodes

**Juan Gonzalez**, Eun Kyung Lee, I-Hsin Chung
IBM Research - High Performance Systems Software

# Motivation and objectives

- **Understand interactions among hardware subsystems** from the performance perspective
  - System usage
  - Application characterization
- IBM integrated profiling for current hybrid architectures (Firestone, Garrison, etc.)
- Integrate existing profiling techniques for CPU + GPU + MPI + …
- Understand interactions among hardware subsystems
  - **To explore optimization opportunities** for applications to achieve greater performance
  - **To analyze hardware performance** to verify/assist architectural design on hybrid nodes

# IBM next-gen profiling: modules & interactions



# Use Case: LULESH CUDA

- The LULESH benchmark represents workloads found in Lagrangian hydrodynamics codes
  - Approximates the hydrodynamics equations discretely by partitioning the problem domain into a collection of volumetric elements defined by a mesh.
  - Designed to support an unstructured mesh with hexahedral elements

# Use Case: LULESH CUDA experimental setup

- 5 Hardware counters groups (22 counters)
- Execution with 8 MPI ranks, 1 CPU per rank, 1 GPU for all ranks.
- Problem size = 100, Iterations = 100



# Use Case: LULESH CUDA `hpm_job_summary`

```
================================================================
Hardware counter summary for job 4746, counter group 13.
Number of MPI ranks in the reporting group = 2.
Max number of MPI ranks per node = 8.
================================================================
----------------------------------------------------------------
mpiAll: call count = 1, avg time (us) = 66404706, max time (us) = 68142052 :
  -- Counter values for processes in this reporting group ----
    min-value  min-rank    max-value  max-rank    avg-value   label
8.224742e+10         5   9.253404e+10         0   8.739073e+10 (PM_CYC) Processor cycles [shared core]
4.321196e+10         5   5.295399e+10         0   4.808297e+10 (PM_CMPLU_STALL) Stalled cycles
7.247977e+09         0   8.297821e+09         5   7.772899e+09 (PM_CMPLU_STALL_THRD) Thread Blocked
9.169238e+10         5   9.464965e+10         0   9.317102e+10 (PM_RUN_INST_CMPL) Instructions completed
4.007753e+08         5   8.858453e+08         0   6.433103e+08 (PM_CMPLU_STALL_BRU_CRU) Stall due to BR or CR
8.224742e+10         5   9.253404e+10         0   8.739073e+10 (PM_RUN_CYC) Run cycles
1.048000e+03         5   1.148000e+03         0   1.098000e+03 MPI p2p communication count
8.863595e+07         0   8.863595e+07         0   8.863595e+07 MPI p2p communication data size
7.661505e+06         0   8.940850e+06         5   8.301178e+06 MPI p2p communication time (us)
2.100000e+01         0   2.100000e+01         0   2.100000e+01 MPI collective communication count
1.600000e+02         0   1.600000e+02         0   1.600000e+02 MPI collective communication data size
2.328000e+04         0   3.993870e+05         5   2.113335e+05 MPI collective communication time (us)
0.000000e+00         0   0.000000e+00         0   0.000000e+00 MPI I/O communication count
0.000000e+00         0   0.000000e+00         0   0.000000e+00 MPI I/O communication data size
0.000000e+00         0   0.000000e+00         0   0.000000e+00 MPI I/O communication time (us)
4.956000e+03         5   5.064000e+03         0   5.010000e+03 CUDA total calls to runtime
3.487636e+07         5   3.543401e+07         0   3.515519e+07 CUDA total time in runtime (us)
1.887000e+03         5   2.047000e+03         0   1.967000e+03 CUDA total kernels executed
3.476472e+07         5   3.534324e+07         0   3.505398e+07 CUDA total time in kernels (us)
2.086202e+08         5   2.284816e+08         0   2.185509e+08 CUDA Host-to-Device bytes transferred
5.725858e+07         5   6.228501e+07         0   5.977179e+07 CUDA Device-to-Host bytes transferred
5.496137e+07         0   7.013558e+07         5   6.254847e+07 inst_executed
8.775782e+07         0   1.092330e+08         5   9.849542e+07 active_cycles
1.597370e+05         0   1.877000e+05         5   1.737185e+05 warps_launched
```

**CPU**

**MPI**

**GPU**

# Use Case: LULESH CUDA `hpm_job_summary`

```
 ------------------------------------------------------------
 CalcKinematicsAndMonotonicQGradient call count = 20, avg time (us) = 321498, max time (us) = 367254 :
   -- Counter values for processes in this reporting group ----
     min-value  min-rank    max-value  max-rank    avg-value   label
   4.247995e+06       5   4.269374e+06       0   4.258684e+06 (PM_CYC) Processor cycles [shared core]
   2.137681e+06       0   2.170562e+06       5   2.154122e+06 (PM_CMPLU_STALL) Stalled cycles
   4.096230e+05       0   4.116920e+05       5   4.106575e+05 (PM_CMPLU_STALL_THRD) Thread Blocked
   2.903952e+06       5   2.904712e+06       0   2.904332e+06 (PM_RUN_INST_CMPL) Instructions completed
   1.057470e+05       0   1.061930e+05       5   1.059700e+05 (PM_CMPLU_STALL_BRU_CRU) Stall due to BR or CR
   4.247995e+06       5   4.269374e+06       0   4.258684e+06 (PM_RUN_CYC) Run cycles
   0.000000e+00       0   0.000000e+00       0   0.000000e+00 MPI p2p communication count
   0.000000e+00       0   0.000000e+00       0   0.000000e+00 MPI p2p communication data size
   0.000000e+00       0   0.000000e+00       0   0.000000e+00 MPI p2p communication time (us)
   0.000000e+00       0   0.000000e+00       0   0.000000e+00 MPI collective communication count
   0.000000e+00       0   0.000000e+00       0   0.000000e+00 MPI collective communication data size
   0.000000e+00       0   0.000000e+00       0   0.000000e+00 MPI collective communication time (us)
   0.000000e+00       0   0.000000e+00       0   0.000000e+00 MPI I/O communication count
   0.000000e+00       0   0.000000e+00       0   0.000000e+00 MPI I/O communication data size
   0.000000e+00       0   0.000000e+00       0   0.000000e+00 MPI I/O communication time (us)
   4.000000e+01       0   4.000000e+01       0   4.000000e+01 CUDA total calls to runtime
   2.751730e+05       0   3.667060e+05       5   3.209395e+05 CUDA total time in runtime (us)
   2.000000e+01       0   2.000000e+01       0   2.000000e+01 CUDA total kernels executed
   2.751630e+05       0   3.666890e+05       5   3.209260e+05 CUDA total time in kernels (us)
   0.000000e+00       0   0.000000e+00       0   0.000000e+00 CUDA Host-to-Device bytes transferred
   0.000000e+00       0   0.000000e+00       0   0.000000e+00 CUDA Device-to-Host bytes transferred
   7.962863e+06       0   1.982396e+07       5   1.389341e+07 inst_executed
   2.349379e+07       0   5.445842e+07       5   3.897611e+07 active_cycles
   5.290000e+03       0   1.308400e+04       5   9.187000e+03 warps_launched
```

# Use Case: LULESH CUDA `hpm_process_summary`

```
=================================================================
Hardware counter data, group = 13, rank = 0.
Number of MPI ranks per node = 8.
=================================================================
-----------------------------------------------------------------
mpiAll  call count = 1, elapsed time = 68142052 (us)
    92534042775 : (PM_CYC) Processor cycles [shared core]
    52953986794 : (PM_CMPLU_STALL) Stalled cycles
     7247976531 : (PM_CMPLU_STALL_THRD) Thread Blocked
    94649648153 : (PM_RUN_INST_CMPL) Instructions completed
      885845339 : (PM_CMPLU_STALL_BRU_CRU) Stall due to BR or CR
    92534042775 : (PM_RUN_CYC) Run cycles
           1148 : MPI p2p communication count
       88635952 : MPI p2p communication data size
        7661505 : MPI p2p communication time (us)
             21 : MPI collective communication count
            160 : MPI collective communication data size
          23280 : MPI collective communication time (us)
              0 : MPI I/O communication count
              0 : MPI I/O communication data size
              0 : MPI I/O communication time (us)
           5064 : CUDA total calls to runtime
       35434011 : CUDA total time in runtime (us)
           2047 : CUDA total kernels executed
       35343244 : CUDA total time in kernels (us)
      228481640 : CUDA Host-to-Device bytes transferred
       62285008 : CUDA Device-to-Host bytes transferred
       54961368 : inst_executed
       87757815 : active_cycles
         159737 : warps_launched
```

# Use Case: LULESH CUDA `hpm_process_summary`

```
-----------------------------------------------------------------
CalcKinematicsAndMonotonicQGradient, call count = 20, elapsed time = 275742 (us)
        4269374 : (PM_CYC) Processor cycles [shared core]
        2137681 : (PM_CMPLU_STALL) Stalled cycles
         409623 : (PM_CMPLU_STALL_THRD) Thread Blocked
        2904712 : (PM_RUN_INST_CMPL) Instructions completed
         105747 : (PM_CMPLU_STALL_BRU_CRU) Stall due to BR or CR
        4269374 : (PM_RUN_CYC) Run cycles
              0 : MPI p2p communication count
              0 : MPI p2p communication data size
              0 : MPI p2p communication time (us)
              0 : MPI collective communication count
              0 : MPI collective communication data size
              0 : MPI collective communication time (us)
              0 : MPI I/O communication count
              0 : MPI I/O communication data size
              0 : MPI I/O communication time (us)
             40 : CUDA total calls to runtime
         275173 : CUDA total time in runtime (us)
             20 : CUDA total kernels executed
         275163 : CUDA total time in kernels (us)
              0 : CUDA Host-to-Device bytes transferred
              0 : CUDA Device-to-Host bytes transferred
        7962863 : inst_executed
       23493788 : active_cycles
           5290 : warps_launched
```

# Use Case: LULESH CUDA `mpi_profile`

```
Data for MPI rank 0 of 8:
Times and statistics from MPI_Init() to MPI_Finalize().
---------------------------------------------------------------
MPI Routine                  #calls     avg. bytes      time(sec)
---------------------------------------------------------------
MPI_Comm_rank                   185            0.0          0.000
MPI_Comm_size                     1            0.0          0.000
MPI_Isend                       207       142428.1          0.011
MPI_Irecv                       347       170470.7          0.001
MPI_Wait                        347            0.0          5.283
MPI_Waitall                      61            0.0          2.367
MPI_Barrier                       1            0.0          0.009
MPI_Reduce                        1            8.0          0.001
MPI_Allreduce                    19            8.0          0.014
---------------------------------------------------------------
MPI task 0 of 8 had the median communication time.
total communication time = 7.685 seconds.
total elapsed time       = 69.496 seconds.
user cpu time            = 28.499 seconds.
system time              = 5.604 seconds.
max resident set size    = 412.500 MBytes.

Rank 3 reported the largest memory utilization : 413.69 MBytes
Rank 0 reported the largest elapsed time : 69.50 sec
```



# Use Case: LULESH CUDA `mpi_profile`

```
Message size distributions:

MPI_Isend                #calls    avg. bytes      time(sec)
                              1           8.0          0.000
                             20          24.0          0.000
                              3         808.0          0.000
                             60        2424.0          0.003
                              3       81608.0          0.000
                            120      242412.0          0.007

MPI_Irecv                #calls    avg. bytes      time(sec)
                              1           8.0          0.000
                             20          24.0          0.000
                             20          48.0          0.000
                              3         808.0          0.000
                             60        2424.0          0.000
                             60        4848.0          0.000
                              3       81608.0          0.000
                            120      242412.0          0.000
                             60      489648.0          0.000

MPI_Reduce               #calls    avg. bytes      time(sec)
                              1           8.0          0.001

MPI_Allreduce            #calls    avg. bytes      time(sec)
                             19           8.0          0.014
```

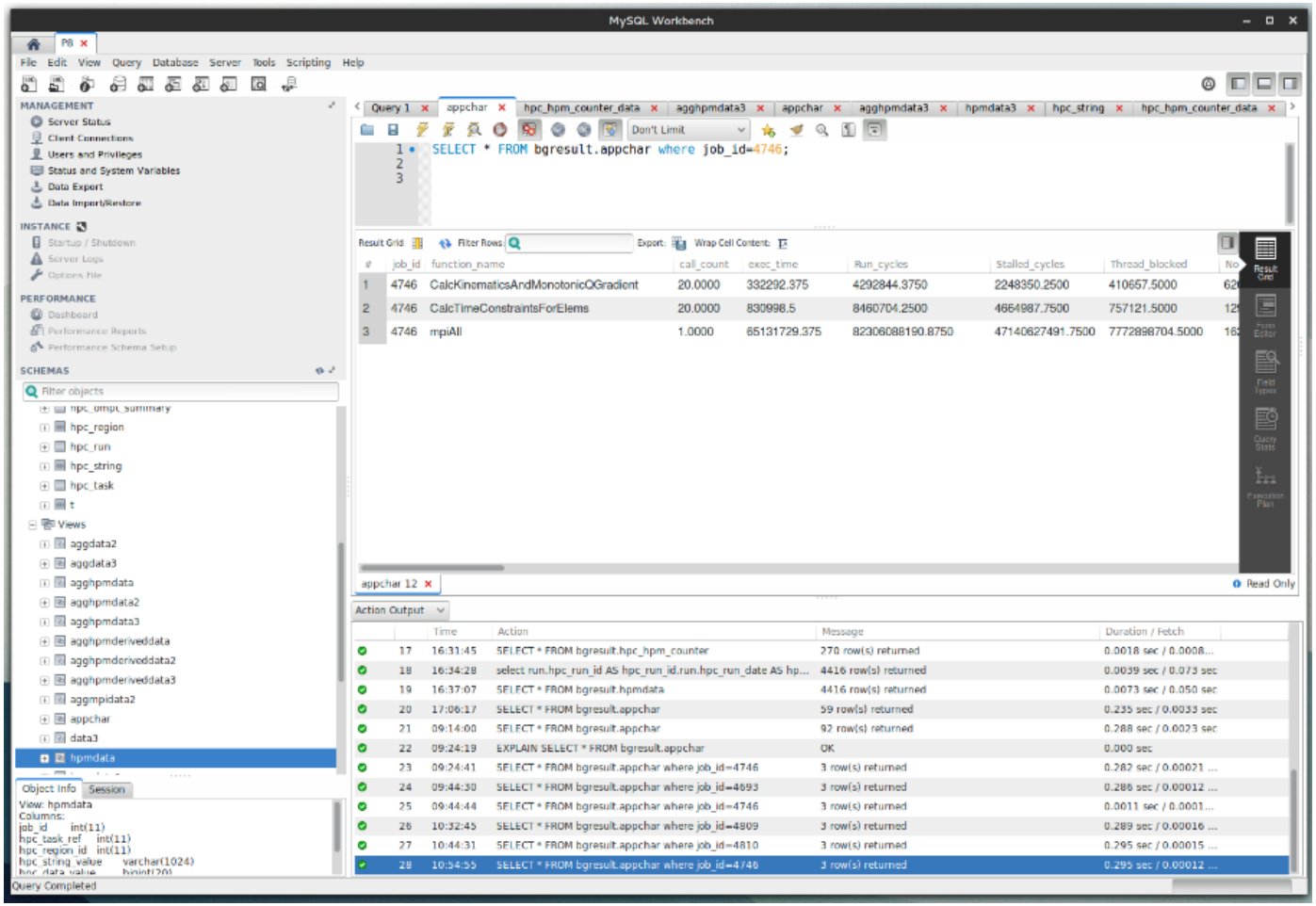
# Use Case: LULESH CUDA `mpi_profile`

```
Communication summary for all tasks:

  minimum communication time = 4.098 sec for task 2
  median  communication time = 7.685 sec for task 0
  maximum communication time = 11.219 sec for task 7


MPI timing summary for all ranks:
taskid      comm(s)  elapsed(s)     user(s)   system(s)    size(MB)   switches
     0         7.68       69.50       28.50        5.60      412.50      51102
     1         8.33       66.03       24.45        5.41      411.12      50155
     2         4.10       66.03       23.58        5.62      411.25      50107
     3         7.48       66.03       23.54        5.53      413.69      50604
     4         7.94       66.03       24.71        5.44      411.31      48061
     5         9.34       66.01       25.35        5.18      412.06      48064
     6         6.60       66.02       25.44        5.51      413.00      47436
     7        11.22       66.00       27.07        5.36      413.25      47294v
```



# Use Case: LULESH CUDA Using the DB

- Structured performance data + standard SQL data access
  - Views to ease data access + GUI (MySQL Workbench)



# Work in Progress

- Work with individual applications
- Support different compilers/ instrumentation methods
- Add more performance data sources (I/O, OpenMP, network, energy)
- Explore modeling with PAPI/native metrics

# Thank you!

# CORAL system profiling

- Sub-routine hooks (TRACE)
  - GCC '-finstrument-function'
- HPM (CPU)
  - PAPI
- MPI (Comms.)
  - PMPI
- CUDA (GPU)
  - CUPTI (Callback + Event APIs)
- HPM+MPI+X



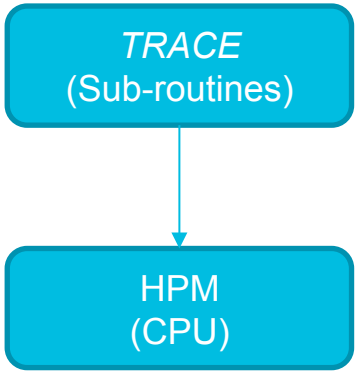
# IBM next-gen profiling: modules & interactions

*TRACE*
(Sub-routines)

- Sub-routine Entry/Exit Hooks + App Init/Fini
  - GCC's –finstrument-function
- Manual instrumentation API
- Based on IBM BG/Q performance repository



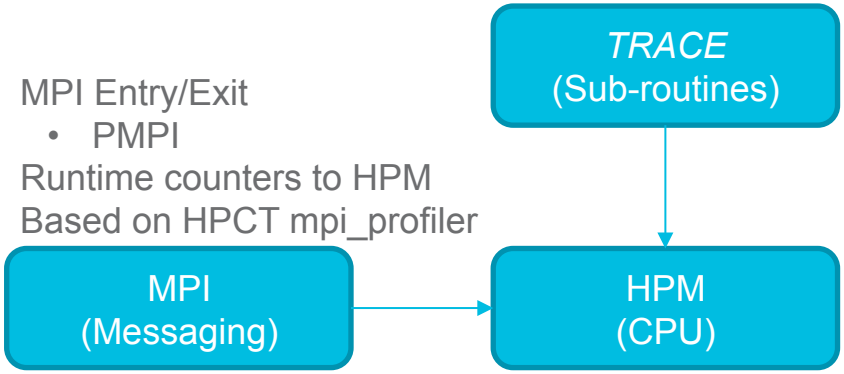
# IBM next-gen profiling: modules & interactions



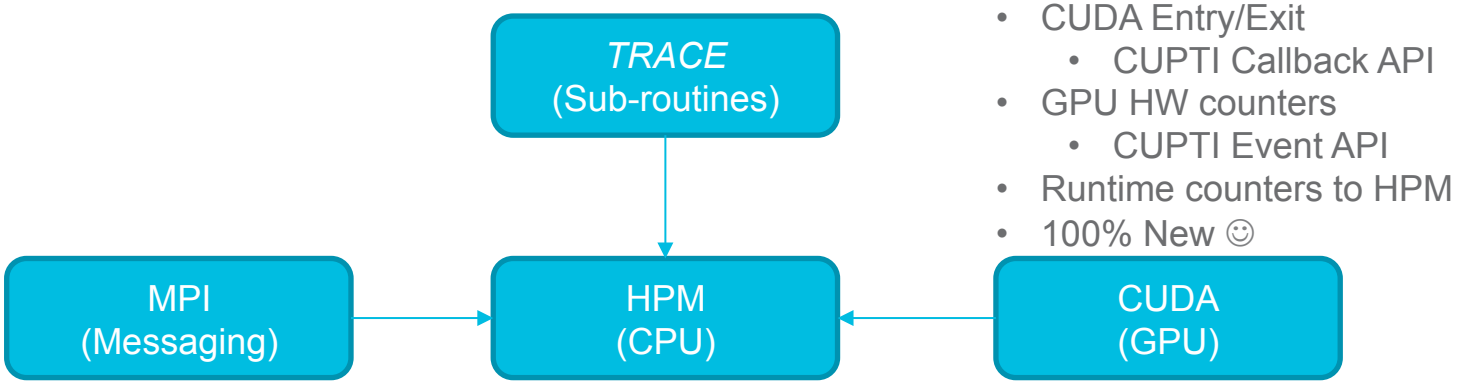

- CPU HW counter measurements
  - PAPI
- Runtime SW counters
- Based on HPCT HPM Code

# IBM next-gen profiling: modules & interactions

- MPI Entry/Exit
  - PMPI
- Runtime counters to HPM
- Based on HPCT mpi_profiler

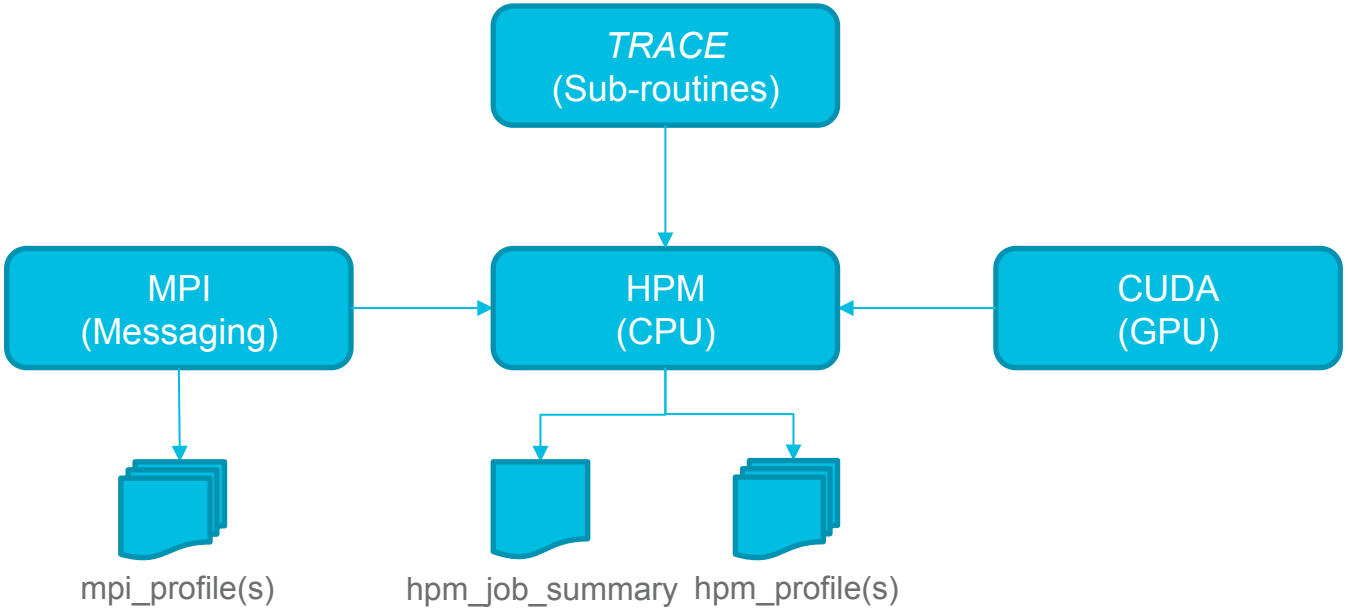


# IBM next-gen profiling: modules & interactions

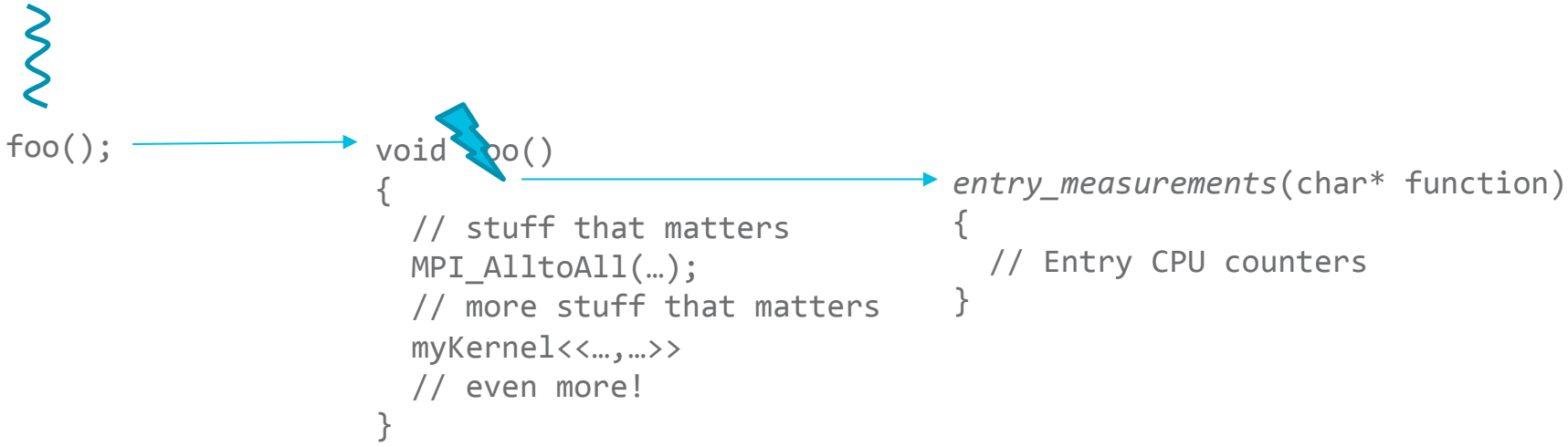



- CUDA Entry/Exit
  - CUPTI Callback API
- GPU HW counters
  - CUPTI Event API
- Runtime counters to HPM
- 100% New ☺

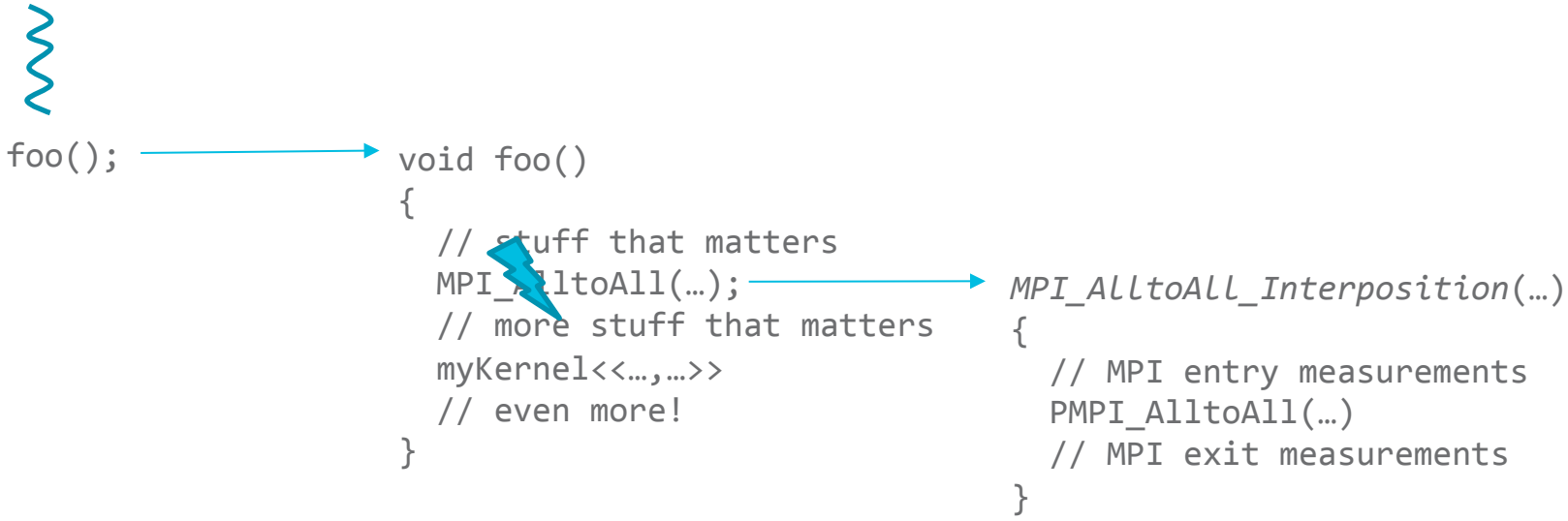
# IBM next-gen profiling: module details



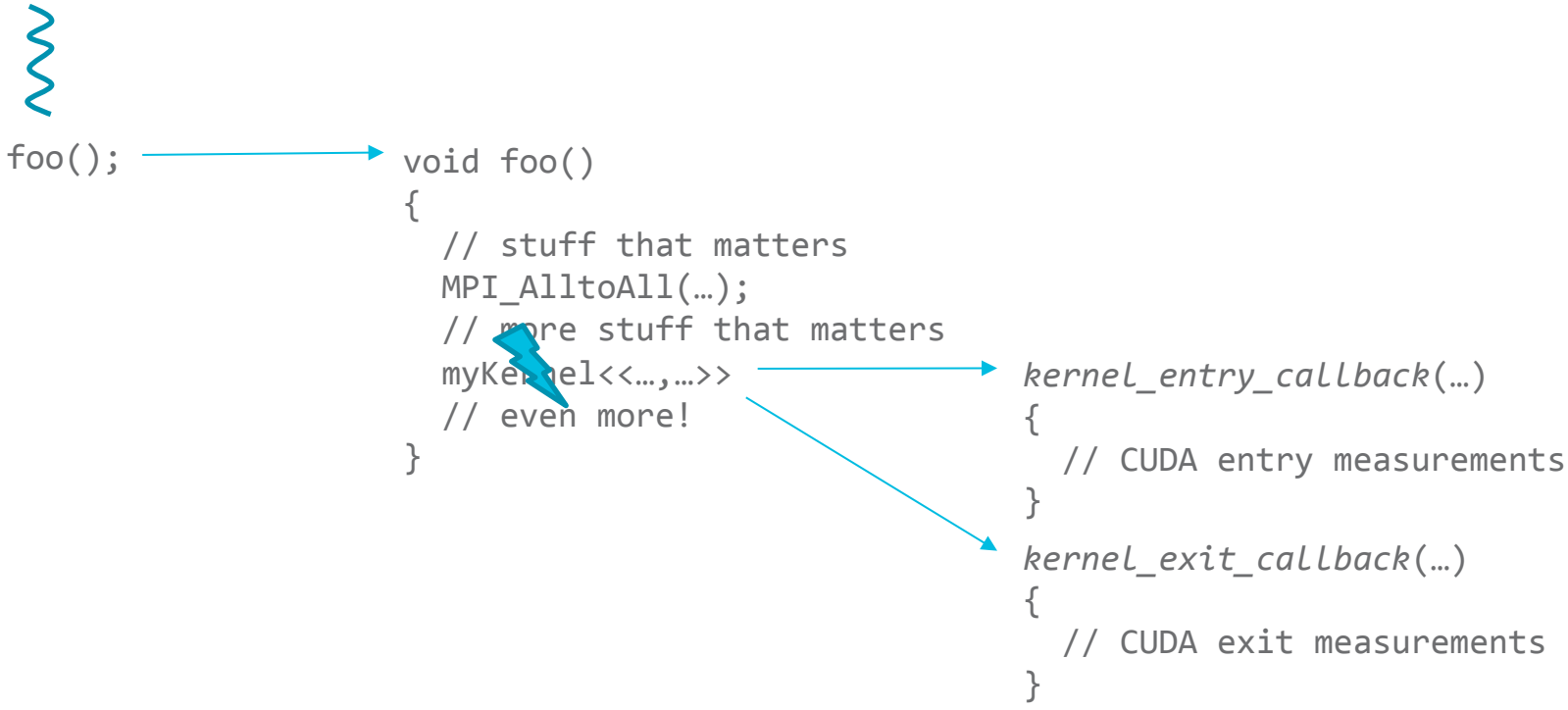
# Usage flow chart



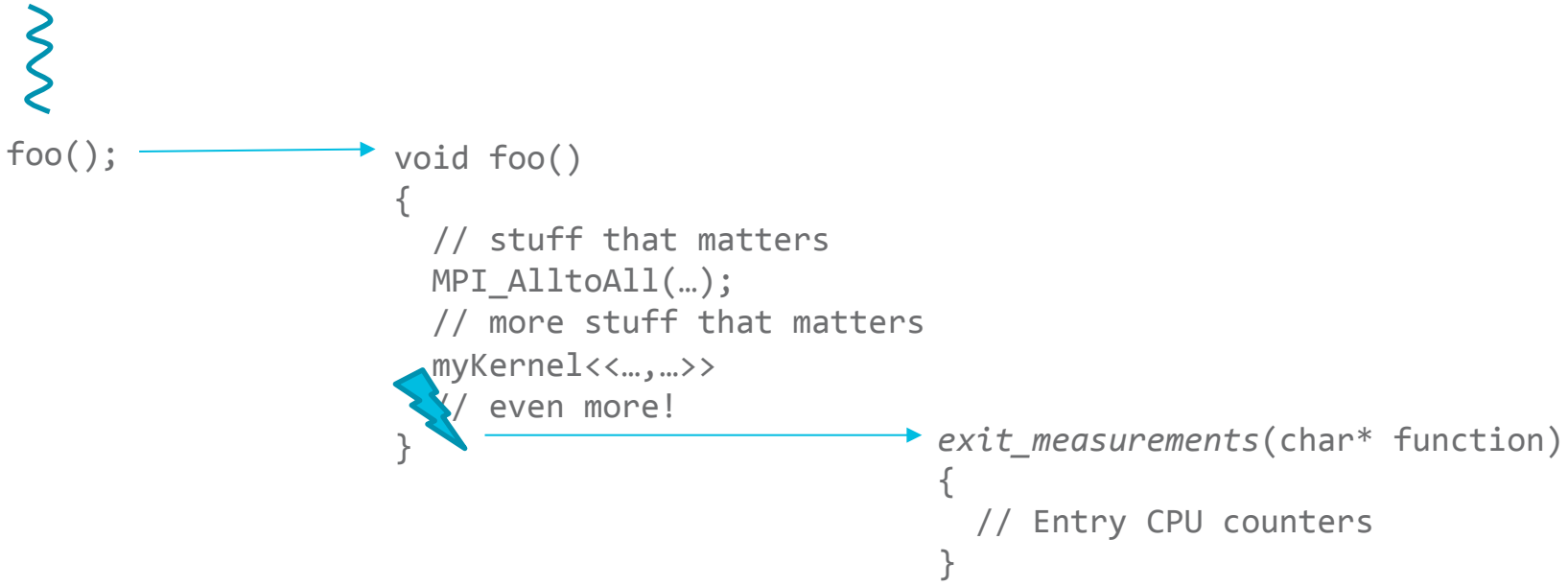
# Usage flow chart



# Usage flow chart



# Usage flow chart



# CUPTI (Profiling GPU)

- Measurements
  - Number of instructions executed
  - Number of warps launched
  - Number of threads launched
  - Memory  usage
  - …..
- Limitation
  - Number of events in a single run
  - Performance degradation while measuring too many events
- Plan: Need to get feedback from the user to collect necessary events